L0: Data Science for Healthcare

Hobson Lane, UC San Diego
Instructor

UC San Diego EXTENSION

UC San Diego SCHOOL OF MEDICINE

Agenda

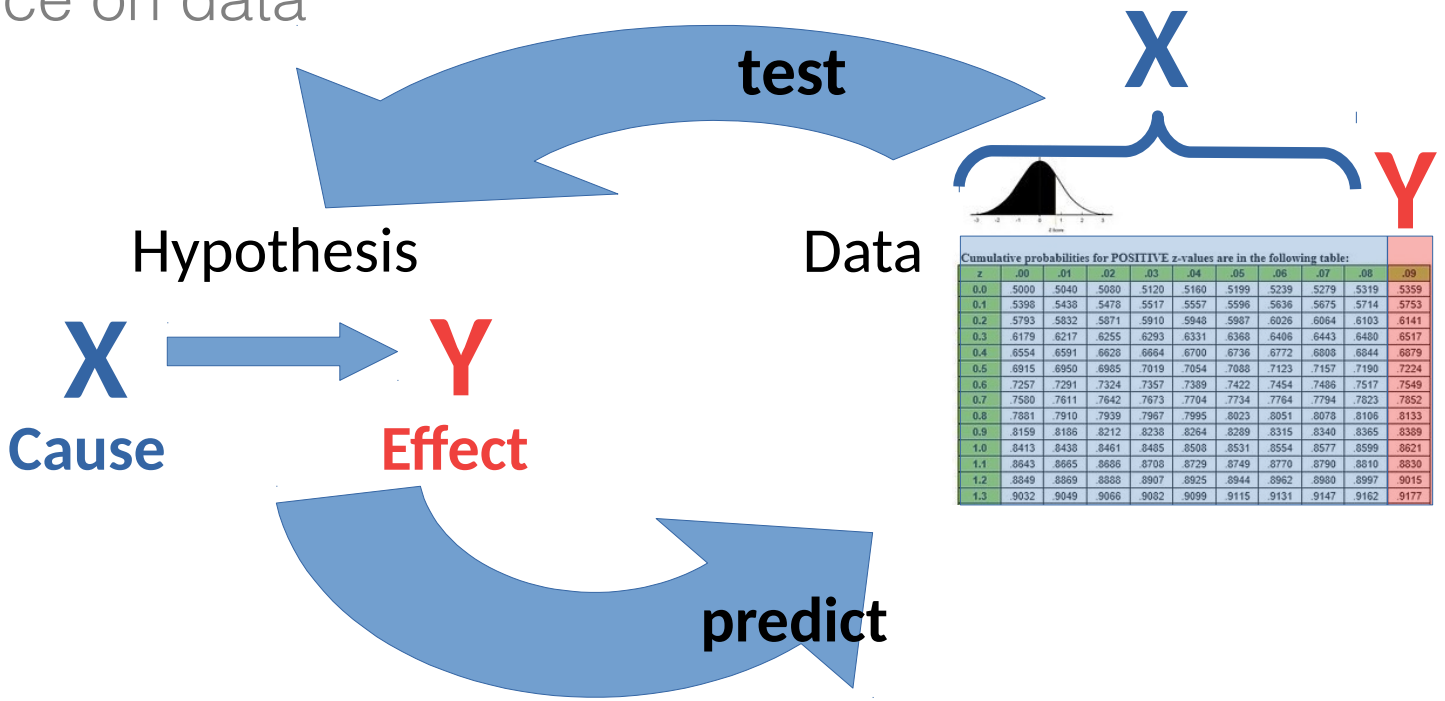
Slide	Topic	Description
3-10	What is Data Science?	Science on data, accidental experiments Trial and error Data Science on “statistics” and “data science” Automation: machine learning and AI
11	Digital Health Data Science course	syllabus
12-14	Example	Kidney Disease, Precision, Recall
15-18	Correlation and causation	Correlation: Mammograms prevent breast cancer death? Spurious correlation “Hill’s Criteria” Causal (influence) diagram
19-21	Bayes Rule	Formula Probabilities for breast cancer and mammograms Mammogram accuracy
22-24	Deep Learning	Multi-layer regression Neural network playground Explaining the black box
25-27	Assignments	Quiz & homework

What is Data Science?

How does it apply to Healthcare?

What is Data Science?

- science on data



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Advantages of the Data Science approach

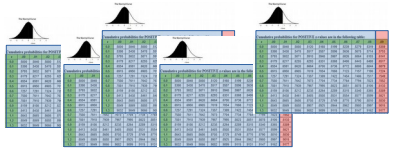
- Accidental experiments are often…

Faster

Cheaper

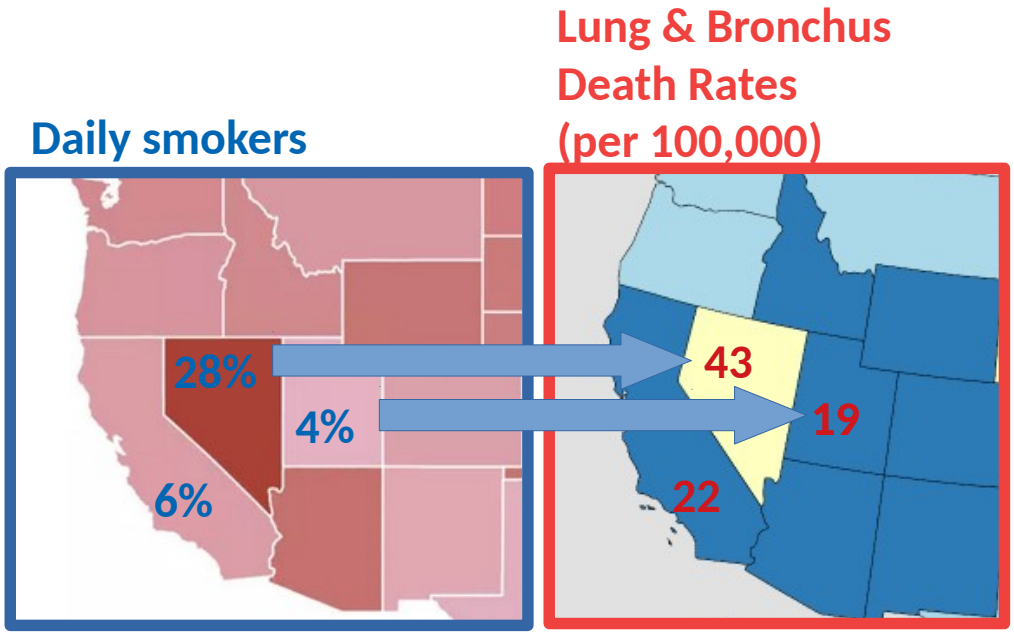
$

Better



More ethical

Smoking → Lung Disease



6	California	22
28	Nevada	43
3	Utah	19

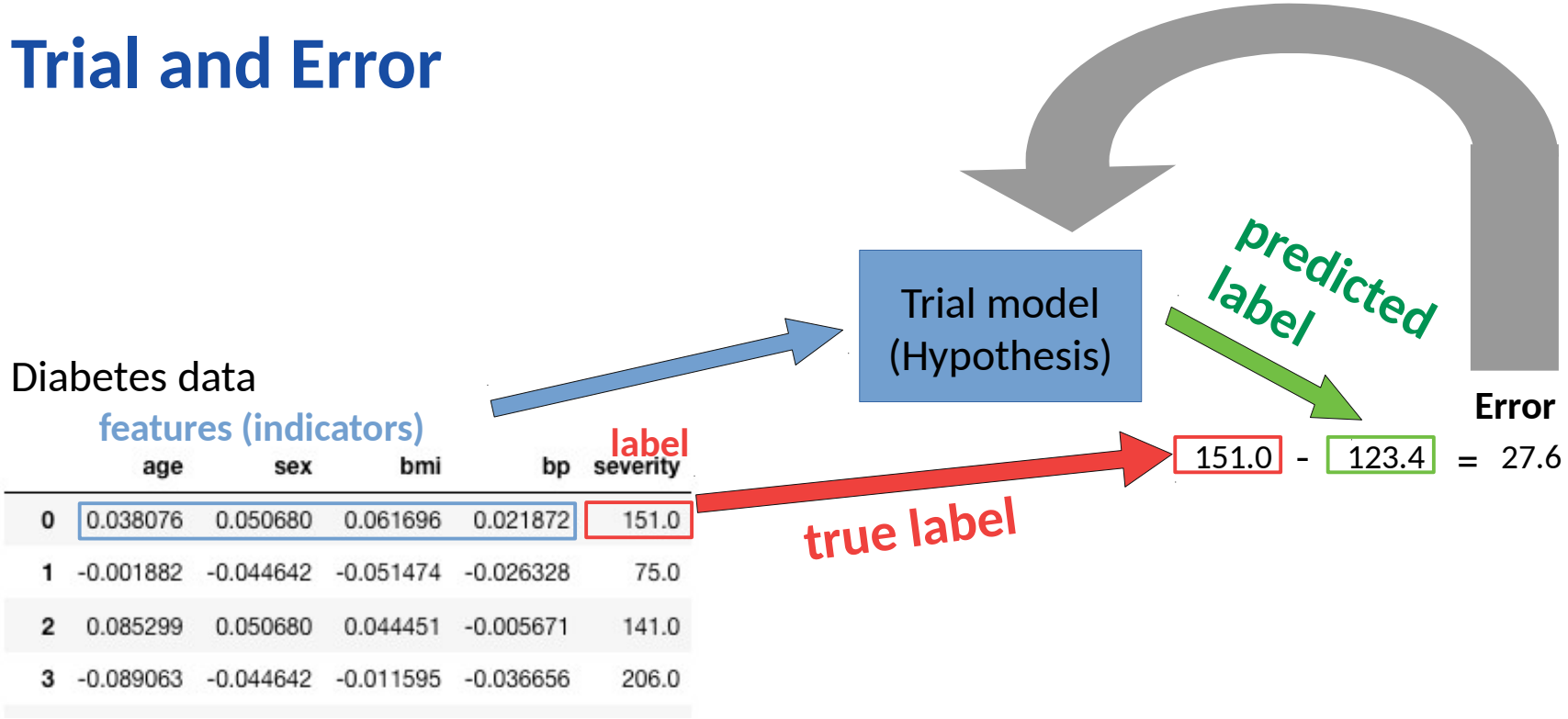
X → **Y**

Smoking **Cancer**

99.98%
Correlation

2% P-value

Trial and Error

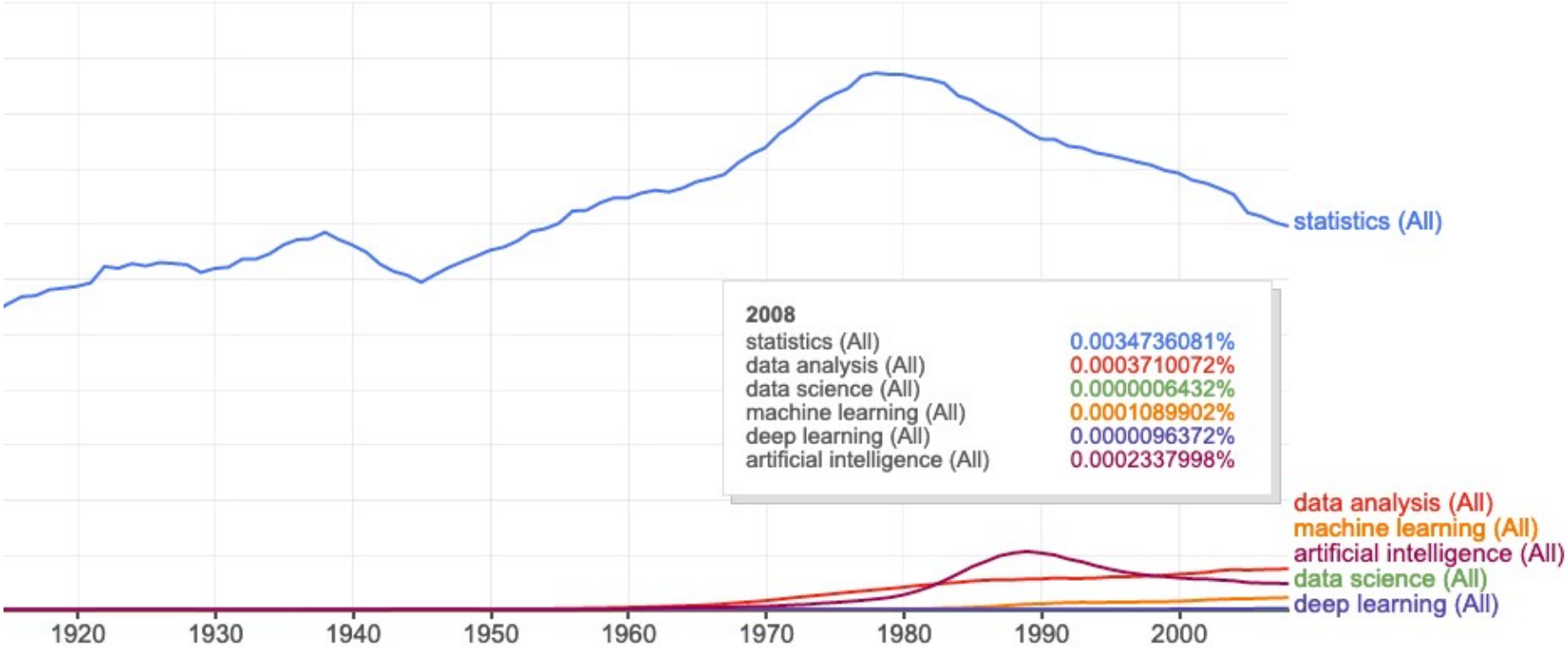


Diabetes data

features (indicators) | label

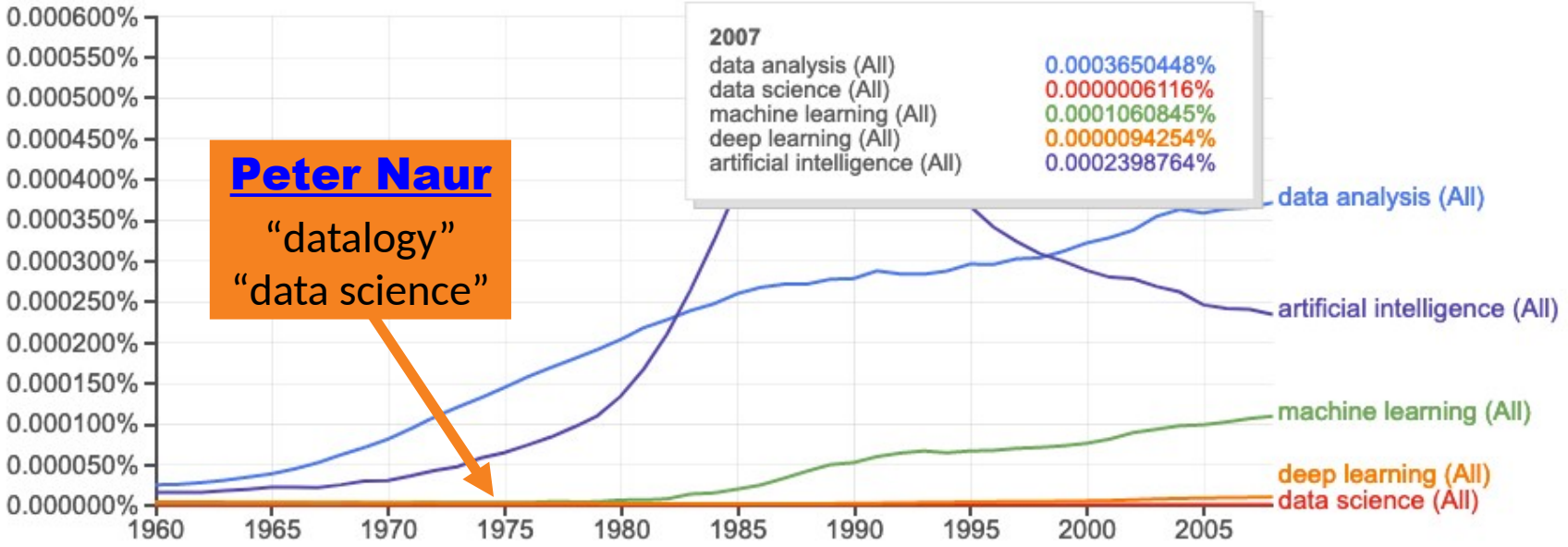
	age	sex	bmi	bp	severity
0	0.038076	0.050680	0.061696	0.021872	151.0
1	-0.001882	-0.044642	-0.051474	-0.026328	75.0
2	0.085299	0.050680	0.044451	-0.005671	141.0
3	-0.089063	-0.044642	-0.011595	-0.036656	206.0

Data science on “statistics”



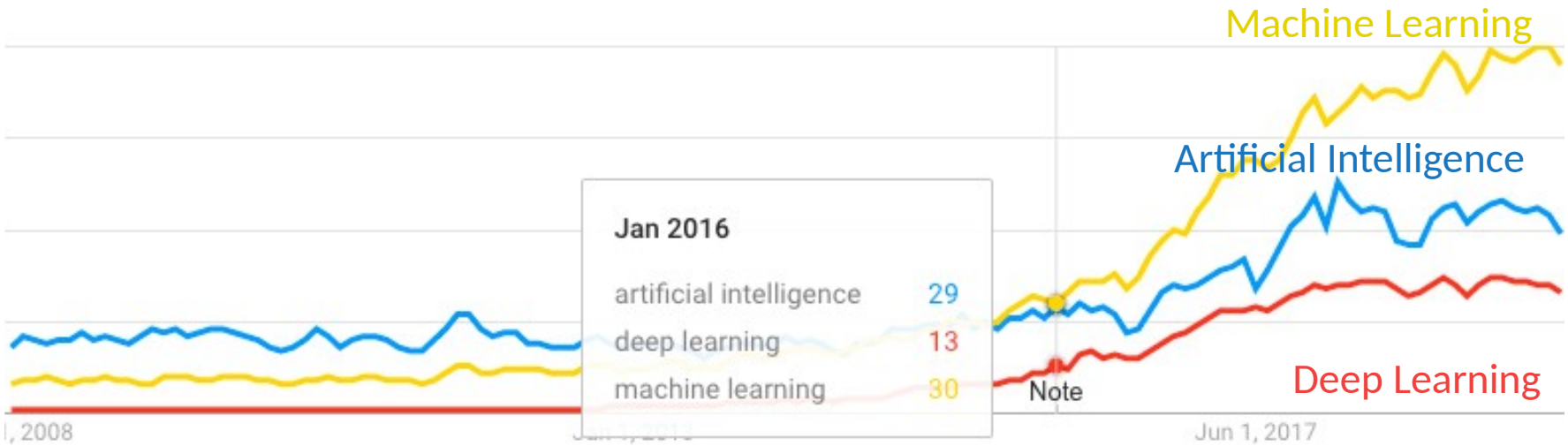
books.google.com/ngrams

Statistics about “data science”



books.google.com/ngrams

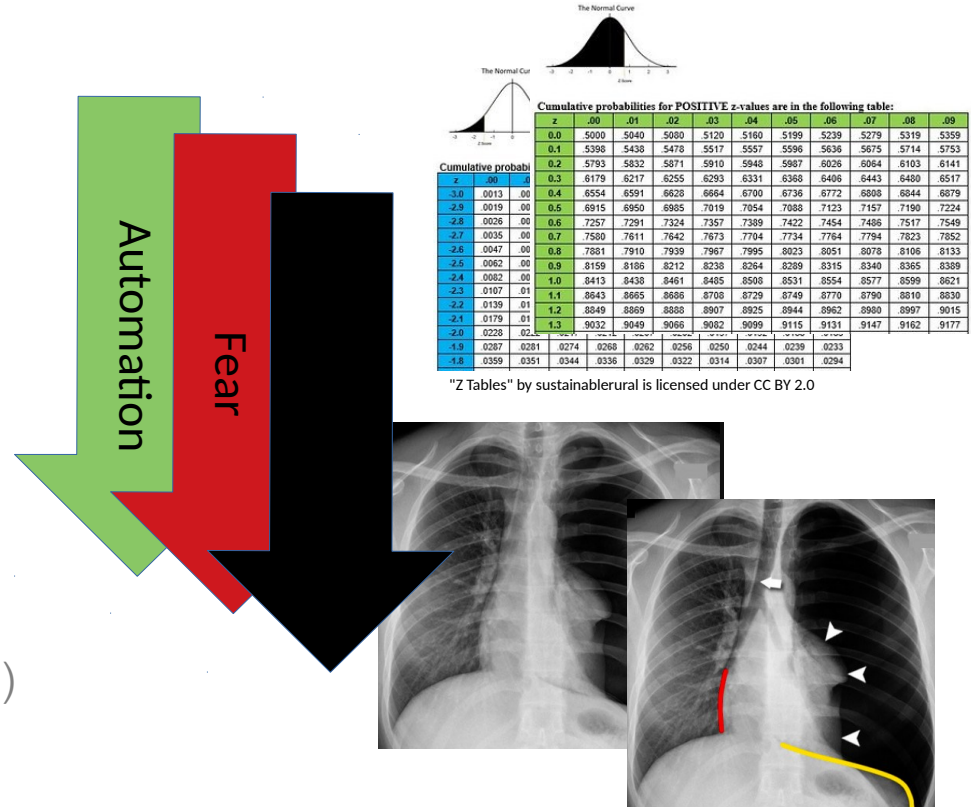
Web search trends



trends.google.com

Automation

- Statistics
- Science
- Data Analytics
- Data Science
- Predictive Analytics
- Machine Learning
- Deep Learning
- Artificial Intelligence (AI)



radiologymasterclass.co.uk

Syllabus

Wk.	Title	Topics	Exercise
1	Data Science in Healthcare	applications, terminology, HIPPA	anonymize dataset PII
2	Spreadsheet Data Science	ETL, exploration & visualization	height, weight, BMI, gender
3	Statistics	causality, correlation, MLE	causal diagram “games”
4	Clinical Data Science & Machine Learning	PII, prescriptive vs descriptive	predict diabetes risk
5	Deep Learning & AI	neural nets, radiology, CV	train diagnostic neural net
6	Hospital Performance Modeling	time series, unintended conseq.	**Mid-term Quiz!**
7	Population Health (Epidemiology)	GIS, spatio-temporal modeling	visualize/analyze Ebola
8	Scoping Review & Gap Analyses	diabetes, military women	review smoking research
9	Natural Language Processing	IA, summarization, text mining	find summary with spaCy
10	Occ. Health & Assistive Tech.	Tesla & Aira case studies, OSHA	analyze medicine OCR app.
11	Public Policy, Privacy and Ethics	bias, fairness, anonymization	**Final Exam!**
Proj	Train your own healthcare model	find/download data, ETL,	fit/train a DS model

Example Application: Predict Kidney Disease

DeepMind (London)

Clinical records can predict Kidney failure

2 days in advance
55% accuracy for acute problems
90% accuracy for serious issues

Dataset:
100% UK citizens
100% military
90% male

Precision

- Of all the positive results how many were correct?
- Positive predictive value
- True positive rate

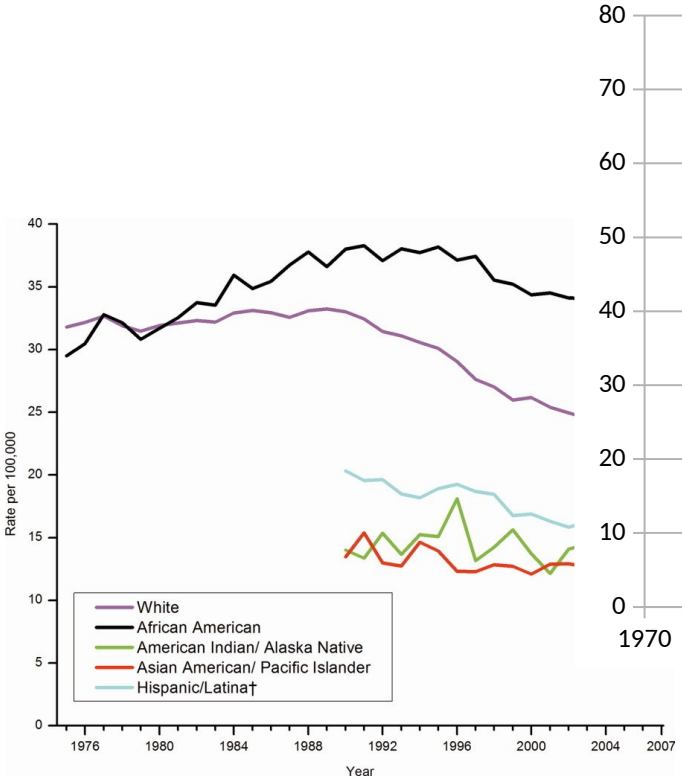
- True_Positive_Count / (True_Positive_Count + False_**Positive**_Count)

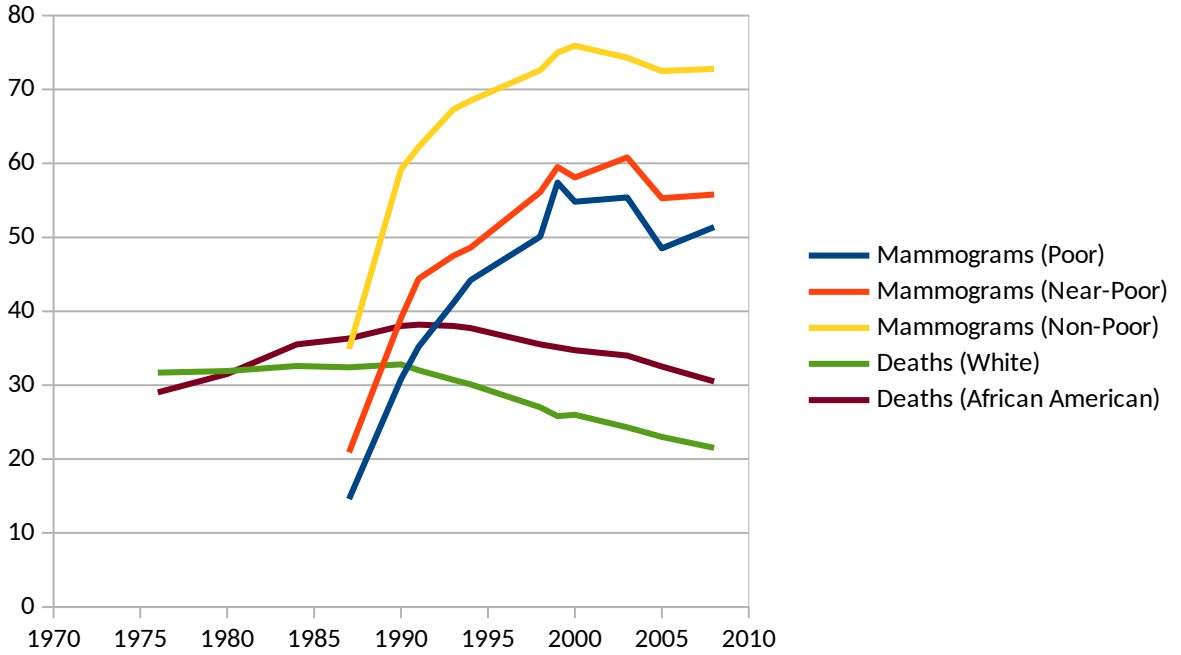
Recall

- Of all the patients with the disease how many were correctly “recalled” (predicted) by the test?
- Sensitivity

- True_Positive_Count / (True_Positive_Count + False_**Negative**_Count)

Correlation enables prediction

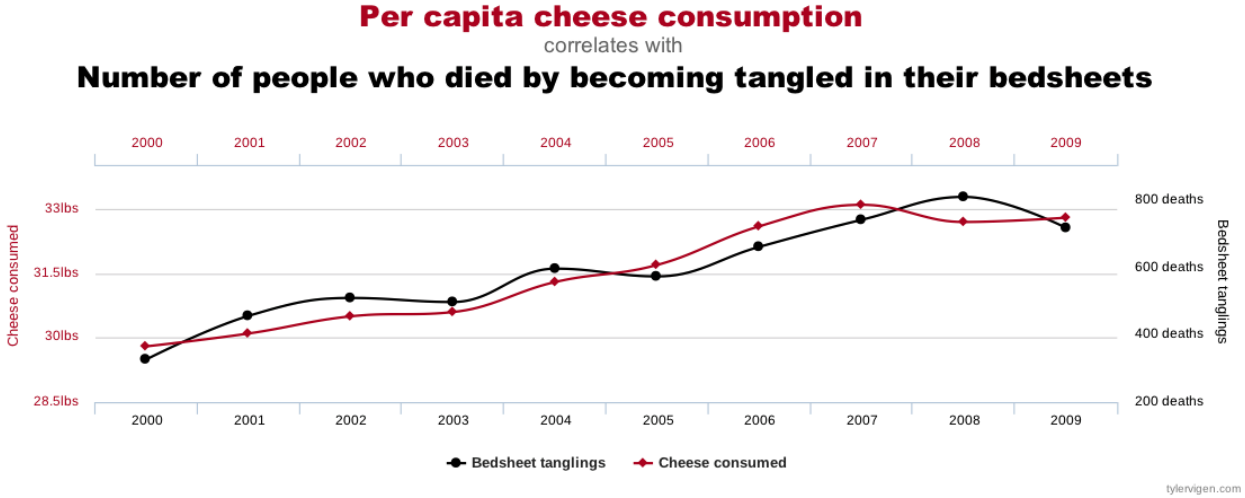




Breast Cancer Rates 2011: bit.ly/ucsdbreast

Correlation is not enough

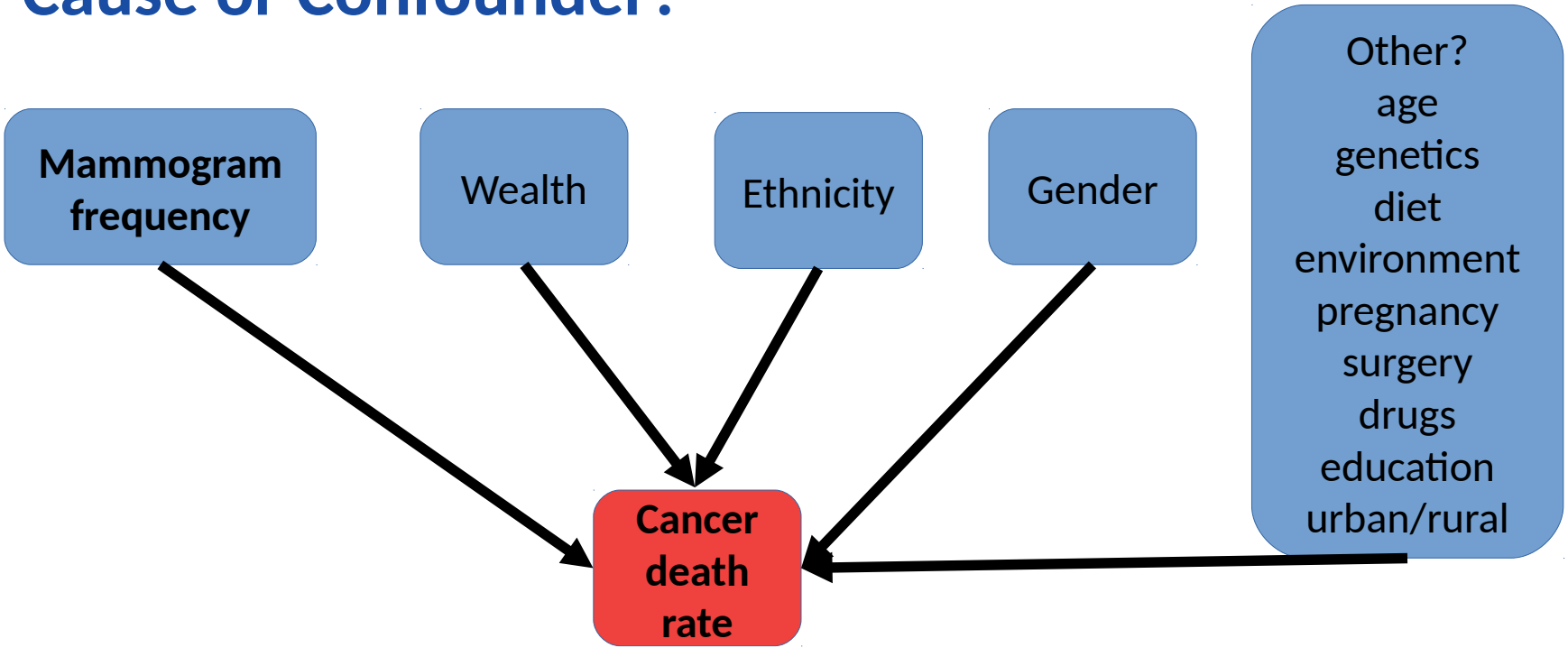
- Computers are good at finding patterns
- But often those patterns are “spurious correlation”



The “Bradford Hill Criteria”

- 1. Consistency: multiple studies, datasets, subjects
- 2. Strength: correlation magnitude
- 3. Dose-response: increased dose = increased effect
- 4. Specificity: 1 effect rather than many
- 5. Temporal relationship: cause before effect
- 6. Coherence: biological plausibility

Cause or Confounder?



Bayes Rule

Updated Probability = Likelihood Ratio X Prior Probability

$$P\left(D|T\right)=\frac{P\left(T|D\right)}{P\left(T\right)}\times P\left(D\right)$$

Bayes Rule Example

Prior	P(D)	Probability of getting breast cancer	1 in 700 per yr 1 in 70,000 (men)
True Positive Rate (Sensitivity)	P(T|D)	Probability of mammogram detecting cancer	.73
False Positive Rate (False Alarm)	P(T|~D)	Probability of positive mammogram w/o cancer	.12
Positive Rate	P(T) = P(D) * P(T|D) + P(~D) * P(T|~D)	Probability of a positive mammogram among all women	.73 * 1 / 700 + .27 * 699 / 700 = .121

Real Numbers

P(D)	1/700
P(T|D)	.73
P(T)	.121

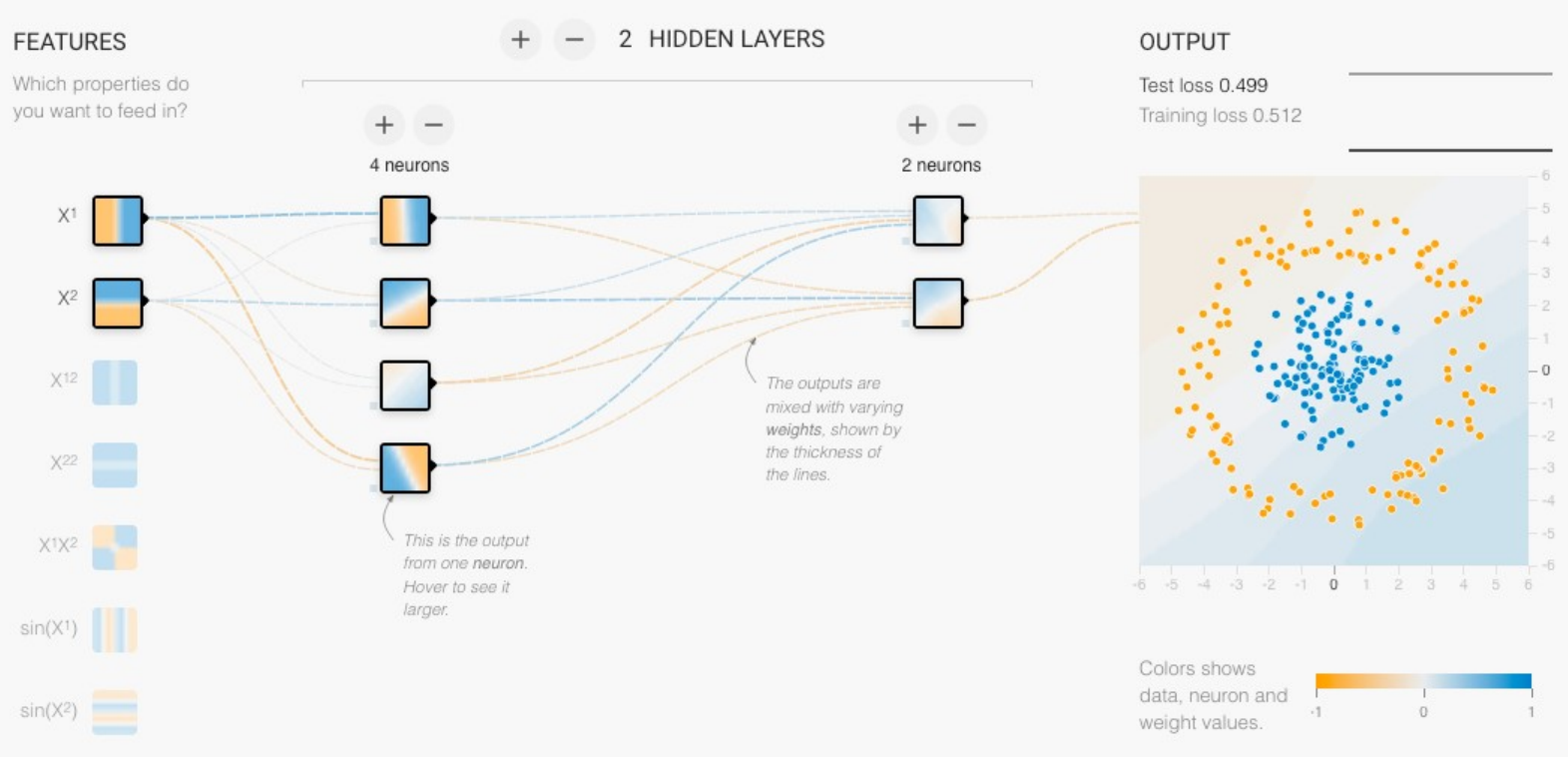
$$P(D|T) = \frac{P(T|D)}{P(T)} \times P(D)$$

$$P(D|T) = \frac{.73}{.121} \times \frac{1}{700} = .0086 \approx 1\%$$

Deep Learning

- Regression works for small numbers of “features”
- Regression can be distracted by spurious correlations
- Feature engineering is the hardest part of Data Science
- What if we layered regressions on top of each other to create a “deeper” model?

Neural Network

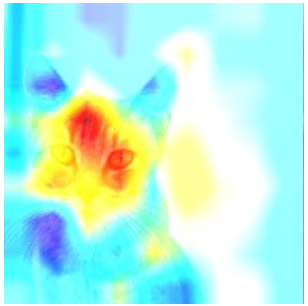


Neural Net Playground: bit.ly/ucsdnet

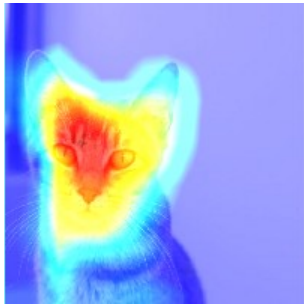
Explainability

Image

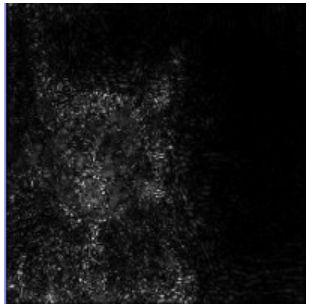
Activations



Importance
(Occlusion Sensitivity)



Grad CAM
(Gradient-weighted Class Activation Map)



SmoothGrad

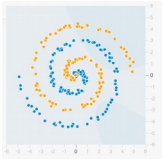
Grad CAM: bit.ly/ucsdcam
tf-explain: bit.ly/ucsdexplain
SmoothGRAD: bit.ly/ucsdsmooth

Assignments

Quiz

1. Give two applications of **Data Science** to Health care
2. How is **Deep Learning** applicable to Health care?
3. Will **Artificial Intelligence** replace doctors?
4. Why or why not?
5. If a blood test for a particular disease has a False Positive rate of 10% and a False Negative rate of 30%, what's the test's *precision* (positive predictive value) and *recall* (sensitivity)?

Homework: Play with Neural Nets



1. Visit playground.tensorflow.org
2. Select the spiral dataset and add 20% Noise
3. Add and remove different combinations of features: x1, x2, x1^2, x2^2, x1*x2, sin(x1), sin(x2)
4. Play around with different numbers of “HIDDEN LAYERS” and neurons per layer.
5. How many features, hidden layers and total neurons do you need achieve < 15% test set loss?